Metabolic quirks and the colourful history of the Euglena gracilis secondary plastid

Novák Vanclová, Anna M. G.; Zoltner, Martin; Kelly, Steven; Soukal, Petr; Záhonová, Kristína; Füssy, Zoltán

Published in:
New Phytologist

DOI:
10.1111/nph.16237

Publication date:
2019

Document Version
Peer reviewed version

MRS ANNA M. G. NOVÁK VANCLOVÁ (Orcid ID : 0000-0002-1613-1625)

Article type : Regular Manuscript

Metabolic quirks and the colourful history of the *Euglena gracilis* secondary plastid

Anna M. G. Novák Vanclová[1], Martin Zoltner[1,2], Steven Kelly[3], Petr Soukal[1], Kristína Záhonová[1,4,5], Zoltán Füssy[5], ThankGod E. Ebenezer[2], Eva Lacová Dobáková[5], Marek Eliáš[4], Julius Lukeš[5,6], Mark C. Field[2,5*] and Vladimír Hampl[1*].

1. Faculty of Science, Charles University, BIOCEV, Vestec, 252 50, Czechia

2. School of Life Sciences, University of Dundee, Dundee, DD1 5EH, UK

3. Department of Plant Sciences, University of Oxford, OX1 3RB Oxford, UK

4. Faculty of Science, University of Ostrava, Ostrava, 701 03, Czechia

5. Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, 370 05, Czechia

6. Department of Biochemistry, University of Cambridge, Cambridge, CB2 1QW, UK

7. Faculty of Science, University of South Bohemia, České Budějovice, 370 05, Czechia

*** Corresponding author contacts:**

–Mark C. Field (proteomics)

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/nph.16237. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

– Vladimír Hampl (plastid evolution): vlada@natur.cuni.cz, +420 325 873 911

Received: 3 June 2019
Accepted: 25 September 2019

Author ORCID numbers:

- Anna M. G. Novák Vanclová: 0000-0002-1613-1625
- Martin Zoltner: 0000-0002-0214-285X
- Steven Kelly: 0000-0001-8583-5362
- Petr Soukal: 0000-0002-3125-253X
- Kristína Záhonová: 0000-0002-5766-0267
- Zoltán Füssy: 0000-0002-0820-6359
- ThankGod E. Ebenezer: 0000-0001-9005-1773
- Eva Lacová Dobáková: N/A
- Marek Eliáš: 0000-0003-0066-6542
- Julius Lukeš: 0000-0002-0578-6618
- Mark C. Field: 0000-0002-4866-2885
- Vladimír Hampl: 0000-0002-5430-7564

Keywords

Euglena gracilis, lateral gene transfer, metabolic reconstruction, plastid, proteome, protein import, shopping bag hypothesis, SUF pathway

Word counts and figure numbers

- Total word count: 6812
- Introduction: 804
- Materials and Methods: 1150
- Results: 3713
- Discussion: 1003
- Acknowledgement: 142

– Number of figures: 7 / 7 in colour

Summary

1. *Euglena* spp. are phototrophic flagellates with considerable ecological presence and impact. *E. gracilis* harbours secondary green plastids, but an incompletely characterized proteome precludes accurate understanding of both plastid function and evolutionary history.
2. Using subcellular fractionation, an improved sequence database and mass spectrometry we determined the composition, evolutionary relationships, and hence predicted functions of the *E. gracilis* plastid proteome.
3. We confidently identified 1,345 distinct plastid protein groups and find that at least 100 proteins represent horizontal acquisitions from organisms other than green algae or prokaryotes. Metabolic reconstruction confirms previously studied/predicted enzymes/pathways and provides evidence for multiple unusual features, including uncoupling of carotenoid and phytol metabolism, a limited role in amino acid metabolism, and dual sets of the SUF pathway for FeS cluster assembly, one of which was acquired by lateral gene transfer from Chlamydiae. Plastid paralogs of trafficking-associated proteins potentially mediating fusion of transport vesicles with the outermost plastid membrane were identified, together with derlin-related proteins, potential translocases across the middle membrane, and an extremely simplified TIC complex.
4. The *Euglena* plastid as the product of many genomes combines novel and conserved features of metabolism and transport.

Introduction

Euglenids are a diverse group of flagellates belonging to the phylum Euglenozoa and have a significant role in the biosphere, as well as being an important model organism and of biotechnological value (Leander *et al.*, 2017). Complex nutritional strategies and an ability to adapt to various environments are major features of the euglenids (Leander *et al.*, 2001; Leander, 2004). The biology of heterotrophic euglenids remains relatively unexplored, but the ease of collection and cultivation of photosynthetic members has made them one of the most widely studied protist groups, despite an absence of a reliable genetic manipulation system. Recent advances in defining the transcriptome and proteome of *Euglena gracilis* promises to improve understanding of the molecular mechanisms employed by this organism and its relatives (Ebenezer *et al.*, 2019).

Euglenophytes harbour green, triple membrane-bound plastids that evolved ~500 Mya from an endosymbiont related to Pyramimonadales (Turmel *et al.*, 2009; Jackson *et al.*, 2018). However, it is possible that other symbionts shared both environment and genes with the ancestors of euglenophytes, as a significant proportion of euglenophyte genes is potentially derived from other algal groups (Maruyama et al. 2011; Markunas and Triemer 2016; Lakey and Triemer 2017; Ponce-Toledo et al. 2018, Ebenezer et al., 2019), possibly originating during multiple rounds of endosymbiotic gene transfer (Larkum *et al.*, 2007). Higher-order endosymbiotic relationships are the result of complex events that can improve our understanding of organellogenesis. Transfer of genes from endosymbiont to host and establishment of new routes for targeting proteins to the organelle are essential for endosymbiont-to-organelle transformation. Protein targeting systems of secondary plastids are notably similar in otherwise unrelated lineages (Durnford & Gray, 2006; Sommer *et al.*, 2007; Hempel *et al.*, 2009; Spork *et al.*, 2009; Minge *et al.*, 2010; Felsner *et al.*, 2011; Lau *et al.*, 2016), suggesting general mechanistic constraints rather than common descent. In brief, components of the secretory pathway have been recruited for protein translocation across the outermost plastid membrane: the process starts as co-translational and signal peptide-dependent import into endoplasmic reticulum (ER) and continues via vesicular transport, either directly or through the Golgi complex (Tonkin *et al.*, 2006; Stiller *et al.*, 2014; Maier *et al.*, 2015), while the major TIC/TOC complex translocases of primary plastids retain their functions at the inner two membranes (van Dooren and Striepen 2013; Sheiner and Striepen 2013; Archibald 2015; Gould et al. 2015; Maier et al. 2015; Bölter and Soll 2016). In plastids of

"chromalveolates", a duplicated version of the ER-associated protein degradation (ERAD) pathway mediates transport across the second-outermost membranes (Spork *et al.*, 2009; Stork *et al.*, 2012; Maier *et al.*, 2015).

In euglenophytes, vesicular transport pathways between the ER, Golgi and plastids are present (Sulli *et al.*, 1999) and have been reconstituted *in vitro* and biochemically suggest the involvement of GTPases but not SNARE proteins (Sláviková *et al.*, 2005), however, the molecular machinery is uncharacterized. A plant-like plastid targeting signal (transit peptide) is essential for plastid import, implying the presence of a TIC/TOC-like pathway (Sláviková *et al.*, 2005). However, *in silico* analysis of the *E. gracilis* and *Euglena longa* transcriptomes identified few TIC subunit homologs, with all TOC subunits failing to be identified even by sensitive HMMER-based approach using multiple strategies for profile building (Záhonová et al. 2018, Ebenezer et al., 2019), implying a highly divergent or alternative translocation mechanism.

The ultrastructure, metabolism, plastid and mitochondrial genomes of *E. gracilis* have been studied in great detail (Tessier *et al.*, 1991; Hallick *et al.*, 1993; Muchhal & Schwartzbach, 1994; Jenkins *et al.*, 1995; Doetsch *et al.*, 2001; Geimer *et al.*, 2009; Mateášiková-Kováčová *et al.*, 2012; Kuo *et al.*, 2013; Dobáková *et al.*, 2015; Watanabe *et al.*, 2017; Gumińska *et al.*, 2018) Moreover, three transcriptome datasets have been generated recently (O'Neill *et al.*, 2015; Yoshida *et al.*, 2016; Ebenezer *et al.*, 2019), the latter coupled with a draft genome and proteomics analysis of whole cell lysates. Features specific to *E. gracilis* were uncovered by these studies, which also enabled predictions of plastid protein composition (Záhonová *et al.*, 2018; Ebenezer *et al.*, 2019). However, such predictions rely on both complete understanding of targeting signals, as well as availability of full-length sequence to encompass N-terminal (and possibly other) plastid targeting peptides. As neither of these criteria is currently met, additional subcellular fractionation evidence is essential for validating predictions and to define a robust proteome. Significantly, the sizes of plastid proteomes vary greatly, from over 1,400 proteins in *Arabidopsis thaliana* to under 400 for the *Plasmodium falciparum* apicoplast (Huang *et al.*, 2013; Boucher *et al.*, 2018).

Here we isolated *E. gracilis* plastids and analysed their composition using unbiased mass spectrometry to identify over 1,300 protein groups. We validate many previous predictions, but also uncovered novel metabolic pathways, illuminated targeting mechanisms and identified many

lateral gene transfer events that support the sequential endosymbiosis model, or "shopping bag hypothesis", for endosymbiotic evolution.

Materials and methods

Isolation of plastid fraction: Plastidial, mitochondrial and peroxisomal fractions of *E. gracilis* were prepared by gradient ultracentrifugation based on protocols used previously (Davis & Merrett, 1973; Dobáková *et al.*, 2015) and described in detail in Methods S1 in the supporting information. The resulting gradient and assessment of fraction quality are documented in Notes S1.2-3.

Mass spectrometry-based identification and quantification of proteins: Plastid and mitochondrial fractions were sonicated in NuPAGE LDS sample buffer (Thermo Scientific) containing 2 mM dithiothreitol and separated on a NuPAGE Bis-Tris 4–12% gradient polyacrylamide gel (Thermo Scientific) under reducing conditions. Each lane was divided into eight slices that were excised, destained and subjected to tryptic digestion and reductive alkylation. These fractions were subjected to liquid chromatography tandem mass spectrometry (LC-MS/MS) on an UltiMate 3000 RSLCnano System (Thermo Scientific) coupled to a Q-exactive mass spectrometer (Thermo Scientific) at the Fingerprints Facility of the University of Dundee. Mass spectra were analysed using MaxQuant (version 1.5, Cox & Mann, 2008), using the predicted translated transcriptome (GEFR00000000.1; Ebenezer *et al.*, 2019). Minimum peptide length was set to six amino acids, isoleucine and leucine were considered indistinguishable and false discovery rates (FDR) of 0.01 were calculated at the levels of peptides, proteins, and modification sites based on the number of hits against a reversed sequence database. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via PRIDE partner repository (Perez-Riverol *et al.*, 2019) with the dataset identifier PXD014767. Ratios were calculated from label-free quantification (LFQ) intensities using only peptides uniquely mapped to a given protein. If the identified peptide sequence set of one protein overlapped another protein, these two proteins were assigned to the same protein group. *P* values were calculated applying t-test based statistics in Perseus (Cox & Mann, 2008; Tyanova *et al.*, 2016). 3,736 distinct protein groups were identified and reduced to 2,544 protein groups by rejecting those groups not identified at the peptide level in at least two of three replicates for one organellar fraction. This filtered set includes a cohort of 774 protein groups that were observed in only one organellar fraction (606 in

the plastid fraction) and in order to form ratios, a constant small value (0.01) was added to the average LFQ intensities of each fraction (to avoid division by zero).

The resulting CP/MT ratio reflects the enrichment of protein groups in the plastid compared to the mitochondrial fraction and is the main indicator for confidence in plastid localization of a given protein in this study. For clarity, CP/MT ratios were $\log_{10}$ transformed and values >3 (CP/MT ratio > 1000) indicated as "3+" in all tables and figures, indicating extremely high, or "infinite" enrichment. The purity of the plastid and mitochondrial fractions was assessed based on distribution and relative abundance of marker proteins (Notes S1 and S2). LFQ analysis of the two organellar fractions and whole cell lysate detected 8,216 protein groups and revealed moderate contamination of both mitochondrial and plastid fractions by other cell compartments (Notes S2.1 and S2.2). Comparison of the plastid and mitochondrial fractions suggests a modest level of cross-contamination between these organelles (Note S2.3).

As proteins encoded by the *E. gracilis* plastid genome (Hallick et al., 1993) were absent from the translated transcriptome, the MaxQuant LFQ analysis was repeated using the plastid gene set and the resulting quantifications of an additional 32 plastid encoded proteins included in the plastid candidate dataset (Dataset S1; seqid prefix "NP"). Additionally, we searched the translated transcriptome of Yoshida et al. (2016; GDJR00000000.1); non-redundant identifications are provided in a separate sheet in Dataset S2. While these proteins were not included in the final plastid candidate dataset, we did consider some of them in the reconstructions presented.

Proteome parsing and annotation: Proteins enriched in the plastid fraction were sorted into four categories based on $\log_{10}$CP/MT ratio (0-1, 1-2, 2-3, and 3+), reflecting the robustness/discrimination of plastid targeting and association. All candidates were annotated using BLAST (Altschul *et al.*, 1997) against the NCBI non-redundant protein database (https://www.ncbi.nlm.nih.gov/protein, 08/16 version), and assigned a KO number by KAAS (http://www.genome.jp/tools/kaas/; Moriya et al., 2007). Annotations were validated manually using UniProt (http://www.uniprot.org/) and OrthoFinder (Emms & Kelly, 2015) and corrected as necessary. Proteins with no, very few (less than five), or very low-scoring (E-value > 1×10^{-5}) homologs identifiable by BLAST were additionally probed using HHpred against PDB, COG, ECOD, and Pfam databases (https://toolkit.tuebingen.mpg.de/#/tools/hhpred; Soding et al., 2005) and assigned a tentative annotation. Proteins with KO and/or EC numbers assigned, proteins

of definite functions in certain metabolic pathways or molecular complexes, as well as proteins with less precise but clear predicted function were sorted manually into 18 custom-defined categories based on the KEGG pathway classification (Note S3). 32 proteins were discarded from the preliminary dataset of 1,377 candidates based on a clearly non-plastid function combined with low enrichment ($\log_{10}$ CP/MT ratio in the 0-1 range; Dataset S1), resulting in a proteome of 1,345 protein groups (Dataset S1).

Bipartite plastid targeting sequences were predicted using a combination of SignalP (version 4.1, http://www.cbs.dtu.dk/services/SignalP/; Petersen et al., 2011) and PrediSI (http://www.predisi.de/; Hiller et al., 2004) for prediction of signal peptides, with ChloroP (http://www.cbs.dtu.dk/services/ChloroP/; Emanuelsson et al., 1999) for prediction of chloroplast transit peptides after *in silico* removal of predicted signal peptides at their putative cleavage sites. Proteins were then binned as having a full bipartite signal, signal peptide only, transit peptide only or no plastid targeting signal sequence.

Major metabolic pathways were reconstructed using the KEGG Mapper web tool (https://www.genome.jp/kegg/tool/map_pathway.html) combined with manual curation. Minor pathways with few members present and/or most members with weak evidence for plastid localization as judged by enrichment ($\log_{10}$ CP/MT ratio close to zero), as well as pathways of known non-plastid localization were excluded.

Determination of the evolutionary origins of plastid proteome members: Each putative plastid protein group was used as query for BLAST searches against 208 transcriptome and genome assemblies from eukaryotes, eubacteria and archaebacteria. Phylogenetic trees containing *E. gracilis* and close homologs were constructed and sorted according to *E. gracilis* protein affiliation supported by bootstrap >= 75% using a semi-automatic pipeline. The description of the bioinformatic pipeline, involving existing bioinformatic software (Stamatakis, 2006; Capella-Gutierrez *et al.*, 2009; Katoh & Standley, 2013) and custom scripts, is given in Methods S1.

N-terminal signal sequence prediction: The hydrophobicity and amino acid composition of the N-termini of plastid-targeted proteins were investigated as described in detail in Methods S1. Analysis was restricted to sequences likely not truncated at the N-terminus, that are translated in the cytoplasm and imported into the plastid, i.e. meeting the following criteria: robust plastid

localization ($\log_{10}$ CP/MT >1.0) and/or predicted photosynthetic function and encoded by complete transcripts containing the spliced leader (ATTTTTTTTCG; Tessier et al., 1991). This cohort consisted of 375 sequences, and a second sequence set of the same size but disregarding plastid localization, was randomly selected from the transcriptome as a control. The differences in amino acid composition were analyzed using $\chi 2$ test (R-Core-Team, 2013).

Data availability

All phylogenetic trees calculated are available at: http://www.protistologie.cz/hampllab/data/Novak_Vanclova_trees.zip.

Results

A plastid proteome for E. gracilis*:* The *E. gracilis* plastid proteome determined here includes 1,345 proteins, 48 of which are encoded by the plastid genome (Hallick *et al.*, 1993); 774 (57.2%) could be functionally annotated while 571 (42.5%) have unknown or hypothetical function and 109 of these (8%) have no identifiable homologs in the nr database at NCBI (Dataset S1). To assess the completeness of the current dataset we found that 48/67 (72%) plastid encoded proteins were present. We also identified 74/83 (89%) and 113/131 (86%) high-confidence plastid-targeted candidates from two earlier studies (Durnford & Gray, 2006; Záhonová *et al.*, 2018; Dataset S2), indicating concordance approaching 90%. By contrast, only 474/1902 (25%) proteins *in silico* predicted as plastid-targeted from transcriptome data (Ebenezer *et al.*, 2019) were detected by proteomics, and 871 proteins identified by proteomics were absent from this *in silico* set, suggesting that available prediction tools are inaccurate in the case of *Euglena*. This could arise partially as a result of N-terminal sequence truncation but could also signify our incomplete understanding of euglenid plastid-targeting signals. It is also relevant that the plastids analysed here were from light grown cells only, and previous work has demonstrated that significant changes to protein composition are not accompanied by transcript changes: some proteins are therefore likely absent from the present plastid material (Ebenezer *et al.*, 2019). It is also possible, that some plastidial proteins are expressed but their physiological abundance or number of ionizable peptides they produce is too low.

Significantly, this new proteome is of similar size to the *Arabidopsis thaliana* plastid proteome (1,462 protein groups; Huang *et al.*, 2013) which has itself been determined using a

combination of *in silico* prediction and proteomic analysis, but is substantially larger than proteomes of plastid or plastid-derived organelles from several unicellular organisms, including *Chlamydomonas reinhardtii* (996; Terashima *et al.*, 2011), *Bigelowiella natans* (324; Hopkins *et al.*, 2012) and the secondarily non-photosynthetic parasite *Plasmodium falciparum* (345; Boucher *et al.*, 2018); see Table 1 for comparison.

Biosynthetic pathways embedded in the Euglena *plastid:* Current state of knowledge regarding metabolic pathway localization in Euglena is largely based on biochemical experiments and *in silico* prediction (recently reviewed in Inwongwan *et al.*, 2019). In this study, we rely on comprehensive proteomic dataset to reconstruct metabolic pathways localized in *E. gracilis* plastid. The 774 functionally annotated proteins were sorted into 18 categories (Fig. 1), and a map of major metabolic pathways and complexes was reconstructed (see Fig. 2 for overall schematic and Notes S5.1-11 for detailed pathways), and while this includes core conserved pathways, we also found features indicative of considerable novelty. As expected, the photosynthetic apparatus is well represented. It should be noted that the precise identification of light-harvesting proteins is not straightforward as these are encoded and translated as polyproteins and are difficult to distinguish using standard proteomic approach without a specific focus on these proteins (Koziol & Durnford, 2008). In accordance with previous findings (Wildner & Hauska, 1974), no gene or protein for plastocyanin was identified (white circle in e^- transport section of Fig. 2, Note S5.2), so electron transfer between the cytochrome b_6f complex and the photosystem I relies solely on two isoforms of cytochrome c_6. Interestingly, *E. gracilis* also encodes a homolog of cytochrome c_{6A} (seqid 17930), previously only known from land plants and green algae (Howe *et al.*, 2006). Cytochrome c_{6A} is thought to function as a redox-sensing regulator within the thylakoid lumen and the structure of its N-terminus in *E. gracilis* is in accord with this localization, but was not detected here due to low abundance and difficulty to detect even in plants (Howe et al., 2006). Notable is also the presence of two homologs of the alpha subunit of the F-type ATPase in the proteome, a canonical copy encoded by the plastid genome and a divergent copy (seqid 3018) encoded in the nucleus. Finally, the *E. gracilis* plastid contains two distinct terminal oxidases (seqids 2887 and 18315), i.e. components of the chloro-respiration pathway (Nawrocki *et al.*, 2015). One represents the conventional enzyme (phylogenetically affiliated with dinoflagellates), whereas the other falls among mitochondrial alternative oxidases (Note S6). The latter is apparently not a contaminant as it is significantly enriched in the plastid fraction and has a

predicted plastid-targeting signal, while at least three different isoforms of the mitochondrial alternative oxidase bearing mitochondrial transit peptides are also present in *E. gracilis*.

The Calvin cycle is fully represented, but a route to glucose metabolism is missing due to the lack of glucose-6-phosphate isomerase. This is probably related to the absence of starch synthesis and a switch to the non-plastid β-1,3-glucan paramylon polysaccharide (Barsanti *et al.*, 2001). The key enzyme for paramylon synthesis, glucan synthase-like 2 (GSL2; Tanaka et al., 2017), is absent from the plastid proteome, but its paralog, GSL1 (seqid 4050), was identified with high-confidence. Unfortunately, the substrate specificity and function of GSL1 are unknown. The nearly complete representation of the C5 pathway of tetrapyrrole synthesis, using glutamate as the precursor of the key intermediate aminolevulinate, is in accord with previous evidence (Gomez-Silva *et al.*, 1985; Kořený & Oborník, 2011).

The *E. gracilis* plastid is also a site for biosynthesis of fatty acids and glycerolipids, the latter including the major plastid phospholipid phosphatidylglycerol and three glycolipids, monogalactosyl- and digalactosyl-diacylglycerol (MGDG and DGDG) and sulfoquinovosyl-diacylglycerol (SQDG) (Matson *et al.*, 1970; Blee & Schantz, 1978; Shibata *et al.*, 2018). We reconstructed pathways for synthesizing these compounds, albeit with several enzymes only predicted by the transcriptome. Interestingly, our data suggests that *E. gracilis* lacks a key enzyme for SQDG synthesis, UDP-sulfoquinovose synthase (SQD1), which catalyses formation of UDP-sulfoquinovose from UDP-glucose. However, sulfite, one of the SQD1 substrates, is readily incorporated into SQDG when incubated with isolated *E. gracilis* plastids (Saidha & Schiff, 1989), suggesting the presence of an alternative pathway to UDP-sulfoquinovose, a substrate of the conventional sulfoquinovosyltransferase (SQD2) catalyzying the ultimate step of SQDG formation (Fig. 2).

Most algal and plant plastids support production of various amino acids, some of which (such as phenylalanine, tyrosine, and tryptophan) are produced exclusively in the organelle, but the *Euglena* plastid lacks these pathways and its contribution to amino acid metabolism is very low. Only an incomplete serine synthesis pathway was reconstructed, with no identifiable plastid-targeted phosphoserine aminotransferase. In addition, only phosphoserine phosphatase was identified in the proteome, whereas the plastid localization of the enzyme catalysing the initial step of the pathway (phosphoglycerate dehydrogenase, divided into two contigs, seqids 23072 and

17279) relies on *in silico* prediction. The *E. gracilis* plastid proteome does contain an isoform of cysteine synthase A, but a plastidial version of the enzyme that converts serine to O-acetylserine, the substrate of cysteine synthase A, was not detected, suggesting that O-acetylserine needs to be imported into the organelle. The plastid further harbours serine/glycine hydroxymethyltransferase (SHMT), whose primary role likely is to provide a formyl group for synthesis of formylmethionyl-tRNA for translation initiation in the plastid rather than to make glycine from serine (Note S5.1). Direct BLAST search of the whole transcriptome confirmed that the enzymes of amino acid biosynthesis are indeed present (*E. gracilis* does not require supplementation by any amino acid when grown autotrophically on minimal medium; Hall & Schoenborn, 1939), but their localization is non-plastidial, in accord with the previous *in silico* prediction (Ebenezer *et al.*, 2019). Possible exceptions are shikimate kinase (seqid 11810) (Inwongwan *et al.*, 2019), which exhibits the typical plastidial targeting signal, and the following enzyme in the shikimate pathway 5-enolpyruvylshikimate-3-phosphate synthase, which has been previously shown to localise in the *E. gracilis* plastid (Reinbothe *et al.*, 1994). However, the former was not detected in the plastid fraction, and the latter is absent from the transcriptomes. Regardless, synthesis of aromatic amino acids in *E. gracilis* may be fully secured by a complete shikimate pathway in the cytosol.

In eukaryotes, two pathways are responsible for synthesis of isopentenyl pyrophosphate (IPP), a precursor to steroids and terpenoids; the MVA pathway, localized to the mitochondria/cytosol, and the MEP (DOXP) pathway, compartmentalized into the plastid (Zhao *et al.*, 2013). *E. gracilis* harbours both pathways, and there is experimental evidence for a role for the MEP pathway in carotenoid synthesis (Kim *et al.*, 2004). Our data demonstrate a plastid localization for the MEP pathway, and also identify the enzyme for the synthesis of the carotenoid precursor geranylgeranyl pyrophosphate (GGPP) from IPP and most enzymes required for synthesis of carotenoids previously reported present in *Euglena* (Krinsky & Goldsmith, 1960): antheraxanthin, neoxanthin, β-, γ- and ζ-carotene, retinol, lutein and cryptoxanthin. Our proteomic and transcriptomic data, however, failed to provide evidence for β-carotene ketolase synthesizing two carotenoids, echinenone and canthaxanthin, that have been detected (Krinsky & Goldsmith, 1960), although some of these enzymes are represented in the transcriptome (marked as grey circles in Fig. 2).

Furthermore, the proteome suggests that IPP, made by the MEP pathway, is used for the synthesis of plastoquinone (via solanesyl-PP), as is usual for plastids in general (Lohr *et al.*, 2012).

In contrast, uniquely among plastid-bearing eukaryotes, the MEP pathway of *Euglena* does not provide GGPP for synthesis of phytyl-PP, a precursor of tocopherols (vitamin E), phylloquinone (vitamin K), and chlorophyll (Kim *et al.*, 2004). Our analyses support this earlier experimental finding and provide an insight into the alternative pathway. Phytyl-PP is conventionally synthesised by ChlP, a GGPP reductase conserved between eukaryotic and prokaryotic phototrophic organisms (Heyes & Neil Hunter, 2009). The *E. gracilis* plastid proteome and transcriptome lack this enzyme, explaining why GGPP synthesised by the MEP pathway cannot serve as a phytyl-PP precursor in this organism. An alternative route to phytyl-PP operates in higher plant chloroplasts, which recycles phytol released during chlorophyll degradation and significantly contributes to tocopherol biosynthesis (Vom Dorp *et al.*, 2015). This pathway consists of VTE5 (phytol kinase, seqid 13197) and VTE6 (phytyl phosphate kinase, seqid 14381), both of which are present in the plastid proteome. The proteome also contains a homolog of a recently characterized chlorophyll dephytylase (CLD1, seqid 12254; Note S5.5; Lin et al., 2016), which further suggests that phytol is salvaged from degraded chlorophyll. This obviously cannot be the only means of producing phytyl-PP, as the compound itself is required for chlorophyll synthesis. Interestingly, the non-photosynthetic chlorophyll-lacking *E. longa* retains both VTE5 and VTE6 (GenBank accession numbers GGOE01004182.1 and GGOE01053790.1) and suggests that phytol is a genuine intermediate for *de novo* phytyl-PP in *Euglena*. As direct phytol synthesis has not been characterized in any organism, how and in which cellular compartment phytol is produced in *Euglena* is thus unclear.

E. gracilis can produce tocopherol (vitamin E) (Watanabe *et al.*, 2017) and homologs of several enzymes for tocopherol synthesis have been described in the transcriptome (O'Neill *et al.*, 2015). There is some uncertainty concerning the subcellular localization of *E. gracilis* tocopherol production (Watanabe *et al.*, 2017), which we now resolve by reconstructing the full pathway, from the precursors homogentisate and phytyl-PP to α-tocopherol, by identification of all four enzymes in the plastid (Fig. 2). Another terpenoid-quinone produced by *E. gracilis* is the phylloquinone derivative 5'-monohydroxyphylloquinone (Ziegler *et al.*, 1989). In plants and green algae most steps of phylloquinone synthesis occur in the plastid, except three reactions from o-succinylbenzoate to dihydroxynaphthoate, which are catalyzed by peroxisomal enzymes (Emonds-Alt *et al.*, 2017; Cenci *et al.*, 2018). *E. gracilis* encodes a homolog of the multifunctional protein PHYLLO catalyzing all four steps of the first part of the pathway (seqid 121), but the protein is

absent from the plastid proteome and lacks a targeting presequence, consistent with biochemical evidence placing synthesis of o-succinylbenzoate in the cytosol (Seeger & Bentley, 1991). The same study tentatively associated the o-succinylbenzoate to dihydroxynaphthoate part of the pathway to the plastid envelope rather than peroxisomes, but the respective enzymes must be unrelated or extremely divergent to the canonical peroxisomal enzymes MenE, MenB, and MenI, as their homologs were not detected in either the plastid proteome or transcriptome. Genome analyses indicate the existence of an novel pathway of dihydroxynaphthoate synthesis in some algae (e.g. glaucophytes) and cyanobacteria (Cenci et al., 2018), so it is possible that *E. gracilis* converts o-succinylbenzoate to dihydroxynaphthoate by employing enzymes of this route. The final part of the pathway, starting with phytylation of dihydroxynaphthoate, is predicted as plastid-localized in *E. gracilis* (Note S5.5), although only one of the respective enzymes (MenA, seqid 7550) was identified in the proteome and the identity of the enzyme catalysing the presumably final step (phylloquinone hydroxylation) is unknown.

Protein import and vesicle targeting to the Euglena *plastid:* A major plastid import pathway is mediated by the TIC/TOC translocation complexes. Surprisingly, only two subunits were identified in the transcriptome: Tic32 and three paralogs of Tic21 (Záhonová *et al.*, 2018). Proteomics confirmed the plastid localization for all Tic21 isoforms, but not Tic32 (Fig. 3). Tic21 is structurally similar to Tic20, the main channel-forming subunit in canonical TIC, and it was sporadically isolated in complex with Tic20 and other central subunits. This suggests that Tic21 is non-essential in plant plastids (Teng *et al.*, 2006; Kikuchi *et al.*, 2009; Nakai, 2018), but its assuming main translocase function in a highly reduced complex is still conceivable. Tic32, on the other hand, is a regulatory subunit with versatile enzymatic activity (Balsera *et al.*, 2010) that could readily serve a TIC-independent purpose or was simply not detected.

Two proteins (seqids 13308 and 11030), representing a novel subgroup of rhomboid-related pseudoproteases distantly related to derlins of the ERAD pathway were also found (Notes S11.1 and S14). In chromalveolate plastids, the ERAD machinery is partially duplicated and recruited for protein transport forming a system termed SELMA, with two derlin family proteins (Der1-1 and Der1-2) presumably constituting the protein-conducting channel (Maier et al. 2015). The *E. gracilis* proteins, however, represent distinct and more distantly related homologs of ERAD derlins. Although the phylogeny is poorly supported it suggests that these proteins

originated by duplication of a host gene, because a third paralogue (seqid 22665, absent from plastid proteome) is present in the transcriptomes of euglenids including heterotroph *Rhabdomonas costata* (Notes S11.2 and S14). It is intriguing to speculate that the detected derlin-like proteins in *E. gracilis* plastid gained a role similar to SELMA through convergent evolution.

The delivery of some plastid-targeted proteins in euglenophytes involves fusion of Golgi-derived vesicles with the outermost plastid membrane and requires GTP hydrolysis, but is not N-ethylmaleimid-sensitive, suggesting the involvement of a Rab/ARF GTPase but not a requirement for conventional SNARE-mediated fusion (Sláviková *et al.*, 2005). Multiple paralogs of some membrane trafficking components are present in *E. gracilis*, and we previously speculated that some of these may be involved in plastid transport (Ebenezer *et al.*, 2019). Interestingly, the plastid proteome contains homologs of coat complex subunits, several Rab GTPases and SNARE proteins, some of which cannot be dismissed as contaminants. For example, SNARE protein GOSR1 has two paralogs in *E. gracilis,* one of which is detected in the plastid proteome with $\log_{10}$ CP/MT ratio > 3 (seqid 18194), and the other (seqid 19152) was not detected; so the apparent insensitivity to NEM might potentially be a technical issue or that GOSR1 functions in an NSF-independent manner. A second notable candidate is Rab5, a conserved Rab GTPase associated with the endosomal system (Langemeyer *et al.*, 2018), again with convincing $\log_{10}$ CP/MT ratio > 3. These two proteins thus may play roles in targeting protein-transporting vesicles to the outermost plastid membrane (Fig. 3), although the current state of understanding of this pathway is rudimentary.

Evolutionary origins of plastid proteins: The semiautomatic phylogenetic pipeline recovered among plastidial proteins 379 (28%) potential lateral gene transfer (LGT) candidates: 346 of these (91%) are related to phototrophic eukaryotes (primary or secondary algae), 20 (5%) to other eukaryotes, and 13 (3%) to prokaryotes (Fig. 4). Most algae-affiliated proteins are related to green algae (95, 28%), as expected given the plastid ancestry. The number of genes likely derived from rhodophytes, cryptophytes, and glaucophytes is very low (around 1-2% each), arguably at the level of technical error. By contrast gene cohorts related to chlorarachniophytes, ochrophytes, and haptophytes (6-7% each, 19% in total) are sufficient to consider as genuine contributions to the *E. gracilis* plastid. Some genes were related to a clan composed of several groups of phototrophs, and therefore assigned to one of several "miscellaneous" categories (Fig. 4). The distribution of the

metabolic functions of proteins affiliated to non-green algae has no clear pattern (Fig. 2). The comparison of these results to the previously published findings on the same topic (Markunas & Triemer, 2016; Ponce-Toledo *et al.*, 2018) is available in Dataset S2.

Additionally, 37 plastidial proteins were determined as affiliated to Discoba, i.e. pre-existing proteins that might have been repurposed or duplicated and recruited for plastidial function. These include mostly membrane transporters, vesicle-associated proteins and chaperones (see Dataset S1). Most of these have $\log_{10}$ CP/MT ratio close to zero and might represent dual-localized proteins or contaminations from other cell fractions, but the functions of the more credible candidates suggest that the host genome contributed to plastid proteome mostly by housekeeping or regulatory functions through ER/Golgi-like proteins and transporters.

Several of the 13 prokaryote-derived proteins encode significant and unexpected functions. For example, four components (SufS, SufB, SufC and SufD) of the SUF system for iron-sulfur (FeS) cluster assembly were found. In transcriptomic data, we also recovered the fifth SUF pathway component, SufE, also of prokaryotic origin. Three of these proteins were specifically related to Chlamydiae in which the SufBCDS genes form a single operon that could have been laterally transferred as one unit from a single source. It is unknown whether these genes cluster together as they were too fragmented in the *E. gracilis* genome assembly (Ebenezer *et al.*, 2019). In general, the SUF pathway endosymbiotically derived from cyanobacteria is essential and universally present in plastids, supplying multiple core enzymes with FeS clusters (Lu, 2018). The chlamydiae-like set of proteins represents a second path for FeS cluster assembly alongside the ancestral plastidial cyanobacteria-related SUF system homologues (Fig. 5, Notes S7 and S8). This apparent redundancy does not represent a transient state and/or result of neutral evolution because these two parallel SUF systems are also present in *E. longa* and *Eutreptiella* spp. (Notes S7 and S8). The relative protein abundance suggests that chlamydiae-like SUF proteins are generally more abundant than the cyanobacteria-like cohort, which suggests that the chlamydiae-like pathway is active and contributes towards plastid FeS metabolism (Fig. 4). Co-occurrence of two SUF pathways in euglenophyte plastids may indicate functional diversification, for example with one in the stroma in common with other plastids (Lu, 2018), and the second restricted to one of the intermembrane spaces, serving co-localized FeS proteins. This hypothesis of spatial separation of the two SUF pathways is further supported by the fact that the chlamydiae-like proteins unlike the other cohort, possess N-terminal extensions which are generally shorter and do not conform to a

typical plastid-targeting domain (Note S15). Similarly, a parallel CIA pathway for FeS cluster assembly was described in cryptophytes and proposed to function in the periplastidial compartment (PPC) which contains a nucleomorph and represents the remnant endosymbiont cytoplasm (Grosche *et al.*, 2018).

A further novel aspect of FeS cluster assembly machinery in *E. gracilis* is the presence of an ABC transporter (seqid 3116) in the plastid proteome, homologous to mitochondrial Atm1 protein required for the export of unspecified FeS cluster intermediates to the cytosol. The presence of this transporter suggests that such intermediates are transported across a membrane, either into the plastid or between sub-compartments (Dataset S1).

While direct experimental data are necessary to understand the function of the second FeS pathway, two additional chlamydial-like proteins were also found, both with orthologs in *E. longa* and *Eutreptiella*. One, an isoform of ferredoxin (divided into two contigs, seqids 37420 and 61063 in our data, seqid 74687 in Yoshida et al. (2016), Dataset S2; Note S9), is itself a FeS protein, and may represent a client of the chlamydial-like SUF. The second is the alpha subunit of NADPH-dependent sulfite reductase (CysJ). Significantly, the beta subunit of this enzyme (CysI) contains a Fe_4S_4 cluster (Smith & Stroupe, 2012) and seems to be related to spirochaetes (Note S10). The plastid localization of sulfite reductase is itself notable, as previous biochemical studies associated sulfate assimilation, including the reaction catalyzed by sulfite reductase, with the mitochondrion in *E. gracilis* (Saidha *et al.*, 1988). The plastid localization of sulfite reduction is consistent with the presence of cysteine synthase in this organelle, as this enzyme assimilates hydrogen sulfide produced in this reaction.

N-terminal plastid-targeting sequences in E. gracilis*:* We took advantage of the confidently identified plastid-targeted proteins to re-evaluate characteristics of *E. gracilis* plastid-targeting sequences. N-terminal sequences of 375 preproteins with well-supported plastid localization ($\log_{10}$ CP/MT ratio > 1 or photosynthetic function) and non-truncated N-termini, verified by the presence of the spliced leader, were analysed. The presence of one or two hydrophobic domains was confirmed using the Kyte-Doolittle hydrophobicity score (Kyte & Doolittle, 1982) in positions 9-60 (putative signal peptide, SP) and 97-130 (putative stop-transfer signal, STS) (Fig. 6). The sequences were sorted into previously defined classes (Durnford & Gray, 2006): 47% class I (SP and STS), 37% class II (SP only), with 16% exhibiting neither of these hydrophobic domains,

referred to as "unclassified". The proportion of "unclassified" preproteins is relatively high and may suggest a significant cohort of plastid proteins with non-conventional targeting signals. No significant correlation between predicted function or preprotein classes was observed.

The canonical euglenophyte plastid-targeting sequence includes a region directly downstream of the SP that exhibits features of a plastid transit peptide (transit peptide-like region, TPL) (Inagaki *et al.*, 2000; Sláviková *et al.*, 2005; Durnford & Gray, 2006), which is surprising, given the apparent absence of a conventional TOC complex in these organisms. To understand the TPL region we investigated its sequence characteristics more closely. The putative TPL region were selected from the 375 plastid proteins, either based on the actual positions of the hydrophobic domains or from TPL region (61-95) estimated from the overall hydrophobicity profile of all studied N-termini. Amino acid compositions of these putative TPLs differed significantly from their respective mature protein regions (Fig. 7). This contrasts to the control sets of 375 randomly selected proteins, where the only significant difference between TPL region and the rest of the protein was a relative enrichment in serine (for full results for both sets and all classes of preproteins, see Notes S12 and S13).

Discussion

The *E. gracilis* plastid proteome described here is similar in size to *Arabidopsis thaliana* (Huang *et al.*, 2013), but larger than *C. reinhardtii* (Terashima *et al.*, 2011) and *B. natans* (Hopkins *et al.*, 2012) suggesting that its functional capacity might be higher than other unicellular phototrophs, which could be connected to the metabolic versatility of euglenophytes. More proteins lacking readily identifiable homologs are present, compared to e.g. *C. reinhardtii,* and the proportion of proteins belonging to equivalent functional categories clearly differs, suggesting divergence in the relative significance of functions. While the former is in part a result of experimental organism bias, this also indicates the importance of analysis of organelles from divergent lineages to gain both an accurate understanding of function in the organism itself as well as across eukaryotic systems.

The euglenid plastid originated from pyramimonadalean green alga (Turmel *et al.*, 2009; Jackson *et al.*, 2018), but a proportion of plastid proteins show distinct phylogenetic affinity. Being noted previously (Maruyama et al. 2011; Markunas and Triemer 2016; Lakey and Triemer

2017; Ponce-Toledo et al. 2018), it was suggested that these genes were acquired from "chromophyte" prey or symbiont by the common ancestor of eukaryovorous and/or phototrophic euglenids. The present set of experimentally determined plastid proteins both supports this hypothesis and allows semiquantitative evaluation of contributions from phototrophic eukaryotes. Gene cohorts related to chlorarachniophytes, ochrophytes and haptophytes account for over 60 *E. gracilis* plastid proteins (19% of the algal LGT candidates). These genes could have originated (1) from the eukaryovorous ancestor of euglenids, *sensu* the "you are what you eat" hypothesis (Doolittle, 1998), (2) from initial stages of endosymbiont integration when the euglenid host was presumably obligatory mixotrophic (much like *Rapaza viridis;* Yamaguchi *et al.*, 2012), such transfers could have compensated for the reductive evolution of the endosymbiont genome, as proposed for the chromatophore of *Paulinella* (Marin *et al.*, 2005; Nowack *et al.*, 2016), (3) gene transfers from a cryptic endosymbiont, kleptoplastid, or even true plastid putatively present in the euglenophyte ancestor and replaced by the extant organelle, in the "shopping bag" (Larkum *et al.*, 2007) or "red carpet" (Ponce-Toledo *et al.*, 2019) sense and similar to the cases of serial endosymbioses and overall plastid fluidity in dinoflagellates (Saldarriaga *et al.*, 2001; Yoon *et al.*, 2005; Takano *et al.*, 2008; Matsumoto *et al.*, 2011; Xia *et al.*, 2013; Burki *et al.*, 2014; Dorrell & Howe, 2015) or (4) horizontal transfers from euglenids to the other algal group. This latter is a possible model for proteins affiliated to chlorarachniophytes given that euglenids are estimated as > 200 MY younger than this group (Jackson et al. 2018).

The proportion of credible plastidial proteins affiliated to Discoba, likely representing ancestral euglenozoan proteins newly recruited for plastidial function, is relatively low. The set is noticeably biased towards proteins related to membrane and vesicular transport, which however, given the methodological constrains of this study, can be interpreted either as the result of the role of host endomembrane system in the transport of biomolecules, or sometimes more obviously as a mere contamination of both analyzed fractions by other subcellular compartments, namely ER, Golgi, and/or vesicles.

The presence of proteins of prokaryotic affiliation is noteworthy and unexpected. Many point at the Chlamydiae as the donor group, the most notable example is a complete second SUF pathway of FeS cluster assembly and hence potentially the result of a single genetic event. A surprisingly high number of chlamydial-like proteins are present in primary plastids of plants and algae (Moustafa *et al.*, 2008; Becker *et al.*, 2008), which led to the "ménage à trois" hypothesis

proposing that a chlamydial endosymbiont was directly implicated in integration of a nascent plastid, and in some manner may have even been essential to the process (Facchinelli *et al.*, 2013; Cenci *et al.*, 2016). Whether this could also be the case for secondary plastid establishment remains an intriguing possibility.

While the *E. gracilis* plastid TPL region is not conserved at the sequence level, there is a characteristic amino acid composition (Fig. 7). In addition to previously described positive net charge, enrichment of hydroxy residues and alanine common with TP of plant plastid proteins (Bruce, 2000; Durnford & Gray, 2006; Patron & Waller, 2007; Felsner *et al.*, 2010; Li & Teng, 2013) there is a paucity of non-polar residues and the high proline content that may confer distinctive secondary structure (MacArthur & Thornton, 1991). Interestingly, proline residues in TP were recently proposed to be important for correct import of plastid proteins with multiple transmembrane domains in plants (Lee *et al.*, 2018), but it is unclear what the significance of general proline-richness in TPs, as observed here, could be. The additional properties of the TPL are likely recognised by an as yet unknown receptor, consistent with a major reorganization to translocation systems, i.e. a reduced TIC/TOC complex and the presence of derlin-like proteins. The absence of unambiguously identifiable additional components, such as Npl4 and Ubx, leaves it unclear if a translocon analogous to SELMA is functional in *E. gracilis*. Nevertheless, we believe it is a noteworthy candidate for components of the unknown euglenophyte plastid protein import system, specifically as translocases of the middle plastid membrane (Fig. 3), filling a gap in the protein import machinery caused by the absence of TOC. These findings also hint at the possibility of the middle membrane being derived from the host endomembrane system and not the outer cyanobacterial-like membrane of the ancestral primary plastid. Most secondary plastids keep all four membranes and it is not clear how probable or improbable the loss of one of the cyanobacterial membranes is, however, the recently described structure of the plastid of the stramenopile *Chrysoparadoxa* suggests that this membrane topology can indeed arise (Wetherbee *et al.*, 2018).

In summary, this is the first report of a proteome of a euglenid plastid, based on direct organellar isolation and proteomics. Novel import systems, metabolic pathways and the presence of significant transfers from prokaryotic genomes all contribute towards a complex and divergent organelle which has a major impact on the biosphere.

Acknowledgements

The authors thank Anna Nenarokova (Biology Centre, České Budějovice) for suggestions and consultations regarding data analysis and Dougie Lamont and colleagues at the Proteomics Facility of the University of Dundee for excellent work. This work was supported by Czech Grant Agency (15-21974S to V.H., 17-21409S to M.E.); ERC CZ (award LL1601 to J.L.); ERD Funds (project OPVVV CZ.02.1.01/0.0/0.0/16_019/0000759 to J.L. and V.H.); Yousef Jameel Academic Program (through the Yousef Jameel PhD Scholarship); the Cambridge Commonwealth; European and International Trust; the Cambridge University Student Registry; the Cambridge Philosophical Society (all to T.E.E.) and the Medical Research Council (Grant #: P009018/1 to M.C.F.). Computational resources were supplied by the Ministry of Education, Youth and Sports of the Czech Republic under the Projects CESNET (Project No. LM2015042) and CERIT-Scientific Cloud (Project No. LM2015085) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Author contributions:

MCF, JL, and VH conceived the original research plans; JL and MCF supervised the experiments; ELD performed the cell fractionation; MZ and SK performed the mass spectrometry analysis; AMGNV and TGE performed protein annotation and sorting, AMGNV, KZ, ZF, and ME interpreted the annotation results, AMGNV performed the signal domain analysis, PS performed the phylogenetic analysis, AMGNV conceived the project and wrote the article with contributions of all the authors; VH, ME, MCF, and JL supervised and complemented the writing. VH agrees to serve as the author responsible for contact and ensuring communication.

References:

Altschul S, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. **1997**. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Archibald JM. **2015**. Genomic perspectives on the birth and spread of plastids. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 1421374112.

Balsera M, Soll J, Buchanan BB. **2010**. Redox extends its regulatory reach to chloroplast protein import. *Trends in plant science* **15**: 515–21.

Barsanti L, Vismara R, Passarelli V, Gualtieri P. **2001**. Paramylon (β-1,3-glucan) content in wild type and WZSL mutant of *Euglena gracilis*. Effects of growth conditions. *Journal of Applied Phycology* **13**: 59–65.

Becker B, Hoef-Emden K, Melkonian M. **2008**. Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evolutionary Biology* **8**: 203.

Blee E, Schantz R. **1978**. Biosynthesis of galactolipids in *Euglena gracilis*: I, Incorporation of UDP galactose into galactosyldiglycerides. *Plant Science Letters* **13**: 247–255.

Bölter B, Soll J. **2016**. Once upon a time - chloroplast protein import research from infancy to future challenges. *Molecular Plant* **9**: 798–812.

Boucher MJ, Ghosh S, Zhang L, Lal A, Jang SW, Ju A, Zhang S, Wang X, Ralph SA, Zou J, *et al.* 2018. Integrative proteomics and bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites (B Striepen, Ed.). *PLOS Biology* **16**: e2005895.

Bruce BD. **2000**. Chloroplast transit peptides: structure, function and evolution. *Trends in Cell Biology* **10**: 440–447.

Burki F, Imanian B, Hehenberger E, Hirakawa Y, Maruyama S, Keeling PJ. **2014**. Endosymbiotic gene transfer in tertiary plastid-containing dinoflagellates. *Eukaryotic Cell* **13**: 246–255.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. **2009**. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Cenci U, Ducatez M, Kadouche D, Colleoni C, Ball SG. **2016**. Was the chlamydial adaptative strategy to tryptophan starvation an early determinant of plastid endosymbiosis? *Frontiers in cellular and infection microbiology* **6**: 67.

Cenci U, Qiu H, Pillonel T, Cardol P, Remacle C, Colleoni C, Kadouche D, Chabi M, Greub G, Bhattacharya D, *et al.* 2018. Host-pathogen biotic interactions shaped vitamin K metabolism in Archaeplastida. *Scientific Reports* **8**: 15243.

Cox J, Mann M. **2008**. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**: 1367–1372.

Davis B, Merrett MJ. **1973**. Malate dehydrogenase isoenzymes in division synchronized cultures of *Euglena*. *Plant Physiology* **51**.

Dobáková E, Flegontov P, Skalický T, Lukeš J. **2015**. Unexpectedly streamlined mitochondrial genome of the euglenozoan *Euglena gracilis*. *Genome Biology and Evolution* **7**: 3358–3367.

Doetsch NA, Favreau MR, Kuscuoglu N, Thompson MD, Hallick RB. **2001**. Chloroplast transformation in *Euglena gracilis*: splicing of a group III twintron transcribed from a transgenic psbK operon. *Current genetics* **39**: 49–60.

Doolittle WF. **1998**. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in genetics : TIG* **14**: 307–11.

van Dooren GG, Striepen B. **2013**. The algal past and parasite present of the apicoplast. *Annual Review of Microbiology* **67**: 271–289.

Vom Dorp K, Hölzl G, Plohmann C, Eisenhut M, Abraham M, Weber APM, Hanson AD, Dörmann P. **2015**. Remobilization of phytol from chlorophyll degradation is essential for tocopherol synthesis and growth of *Arabidopsis*. *The Plant cell* **27**: 2846–59.

Dorrell RG, Howe CJ. **2015**. Integration of plastids with their hosts: Lessons learned from dinoflagellates. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 10247–54.

Durnford DG, Gray MW. **2006**. Analysis of *Euglena gracilis* plastid-targeted proteins reveals different classes of transit sequences. *Eukaryotic Cell* **5**: 2079–2091.

Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novák Vanclová AMG, Prasad B, Soukal P, Santana-Molina C, O'Neill E, Nankissoor NN, *et al.* 2019. Transcriptome, proteome and draft genome of *Euglena gracilis*. *BMC Biology* **17**: 11.

Emanuelsson O, Nielsen H, von Heijne G. **1999**. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science : A Publication of*

the Protein Society **8**: 978–984.

Emms DM, Kelly S. **2015**. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**: 157.

Emonds-Alt B, Coosemans N, Gerards T, Remacle C, Cardol P. **2017**. Isolation and characterization of mutants corresponding to the MENA, MENB, MENC and MENE enzymatic steps of 5'-monohydroxyphylloquinone biosynthesis in *Chlamydomonas reinhardtii*. *The Plant journal : for cell and molecular biology* **89**: 141–154.

Facchinelli F, Colleoni C, Ball SG, Weber APM. **2013**. Chlamydia, cyanobiont, or host: who was on top in the ménage à trois? *Trends in plant science* **18**: 673–9.

Felsner G, Sommer MS, Gruenheit N, Hempel F, Moog D, Zauner S, Martin W, Maier UG. **2011**. ERAD components in organisms with complex red plastids suggest recruitment of a preexisting protein transport pathway for the periplastid membrane. *Genome biology and evolution* **3**: 140–50.

Felsner G, Sommer MS, Maier UG. **2010**. The physical and functional borders of transit peptide-like sequences in secondary endosymbionts. *BMC plant biology* **10**: 223.

Geimer S, Belicová A, Legen J, Sláviková S, Herrmann RG, Krajcovic J. **2009**. Transcriptome analysis of the *Euglena gracilis* plastid chromosome. *Current genetics* **55**: 425–38.

Gomez-Silva B, Timko MP, Schiff JA. **1985**. Chlorophyll biosynthesis from glutamate or 5-aminolevulinate in intact *Euglena* chloroplasts. *Planta* **165**: 12–22.

Gould SB, Maier U-G, Martin WF. **2015**. Protein import and the origin of red complex plastids. *Current biology : CB* **25**: R515–R521.

Grosche C, Diehl A, Rensing SA, Maier UG. **2018**. Iron–sulfur cluster biosynthesis in algae with complex plastids (M Embley, Ed.). *Genome Biology and Evolution* **10**: 2061–2071.

Gumińska N, Płecha M, Zakryś B, Milanowski R. **2018**. Order of removal of conventional and nonconventional introns from nuclear transcripts of *Euglena gracilis* (MC Field, Ed.). *PLOS Genetics* **14**: e1007761.

Hall RP, Schoenborn HW. **1939**. The question of autotrophic nutrition in *Euglena gracilis*.

Physiological Zoology. **12**: 76-84.

Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E. **1993**. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic acids research* **21**: 3537–3544.

Hempel F, Bullmann L, Lau J, Zauner S, Maier UG. **2009**. ERAD-derived preprotein transport across the second outermost plastid membrane of diatoms. *Molecular biology and evolution* **26**: 1781–90.

Heyes DJ, Neil Hunter C. **2009**. Biosynthesis of chlorophyll and bacteriochlorophyll. In: Warren MJ, Smith AG, eds. Tetrapyrroles. New York, NY: Springer New York, 235–249.

Hiller K, Grote A, Scheer M, Münch R, Jahn D. **2004**. PrediSi: Prediction of signal peptides and their cleavage positions. *Nucleic Acids Research* **32**: 375–379.

Hopkins JF, Spencer DF, Laboissiere S, Neilson JAD, Eveleigh RJM, Durnford DG, Gray MW, Archibald JM. **2012**. Proteomics reveals plastid- and periplastid-targeted proteins in the chlorarachniophyte alga *Bigelowiella natans*. *Genome Biology and Evolution* **4**: 1391–1406.

Howe CJ, Schlarb-Ridley BG, Wastl J, Purton S, Bendall DS. **2006**. The novel cytochrome c6 of chloroplasts: a case of evolutionary bricolage? *Journal of Experimental Botany* **57**: 13–22.

Huang M, Friso G, Nishimura K, Qu X, Olinares PDB, Majeran W, Sun Q, van Wijk KJ. **2013**. Construction of plastid reference proteomes for maize and *Arabidopsis* and evaluation of their orthologous relationships; the concept of orthoproteomics. *Journal of Proteome Research* **12**: 491–504.

Inagaki J, Fujita Y, Hase T, Yamamoto Y. **2000**. Protein translocation within chloroplast is similar in *Euglena* and higher plants. *Biochemical and biophysical research communications* **277**: 436–442.

Inwongwan S, Kruger NJ, Ratcliffe RG, O'Neill EC. **2019**. *Euglena* central metabolic pathways and their subcellular locations. *Metabolites*. **9**: 115.

Jackson C, Knoll AH, Chan CX, Verbruggen H. **2018**. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green

plastids. *Scientific Reports* **8**: 1523.

Jenkins KP, Hong L, Hallick RB. **1995**. Alternative splicing of the *Euglena gracilis* chloroplast *roaA* transcript. *RNA (New York, N.Y.)* **1**: 624–633.

Katoh K, Standley DM. **2013**. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kikuchi S, Oishi M, Hirabayashi Y, Lee DW, Hwang I, Nakai M. **2009**. A 1-megadalton translocation complex containing Tic20 and Tic21 mediates chloroplast protein import at the inner envelope membrane. *The Plant cell* **21**: 1781–97.

Kim D, Filtz MR, Proteau PJ. **2004**. The methylerythritol phosphate pathway contributes to carotenoid but not phytol biosynthesis in *Euglena gracilis*. *Journal of Natural Products* **67**: 1067–1069.

Kořený L, Oborník M. **2011**. Sequence evidence for the presence of two tetrapyrrole pathways in *Euglena gracilis*. *Genome Biology and Evolution* **3**: 359–364.

Koziol AG, Durnford DG. **2008**. *Euglena* light-harvesting complexes are encoded by multifarious polyprotein mRNAs that evolve in concert. *Molecular Biology and Evolution* **25**: 92–100.

Krinsky NI, Goldsmith TH. **1960**. The carotenoids of the flagellated alga, *Euglena gracilis*. *Archives of biochemistry and biophysics* **91**: 271–279.

Kuo RC, Zhang H, Zhuang Y, Hannick L, Lin S. **2013**. Transcriptomic study reveals widespread spliced leader trans-splicing, short 5'-UTRs and potential complex carbon fixation mechanisms in the euglenoid alga *Eutreptiella sp*. *PLoS ONE* **8**: e60826.

Kyte J, Doolittle RF. **1982**. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**: 105–132.

Lakey B, Triemer R. **2017**. The tetrapyrrole synthesis pathway as a model of horizontal gene transfer in euglenoids (K Müller, Ed.). *Journal of Phycology* **53**: 198–217.

Langemeyer L, Fröhlich F, Ungermann C. **2018**. Rab GTPase function in endosome and lysosome biogenesis. *Trends in cell biology* **28**: 957–970.

Larkum AWD, Lockhart PJ, Howe CJ. **2007**. Shopping for plastids. *Trends in Plant Science* **12**: 189–195.

Lau JB, Stork S, Moog D, Schulz J, Maier UG. **2016**. Protein-protein interactions indicate composition of a 480 kDa SELMA complex in the second outermost membrane of diatom complex plastids. *Molecular Microbiology* **100**: 76–89.

Leander BS. **2004**. Did trypanosomatid parasites have photosynthetic ancestors? *Trends in Microbiology* **12**: 251–258.

Leander BS, Lax G, Karnkowska A, Simpson AGB. **2017**. Euglenida. In: Archibald J., Simpson A, Slamovits C, eds. Handbook of the Protists. Cham: Springer International Publishing, 1–42.

Leander BS, Triemer RE, Farmer M a. **2001**. Character evolution in heterotrophic euglenids. *European Journal of Protistology* **37**: 337–356.

Lee DW, Yoo Y-J, Razzak MA, Hwang I. **2018**. Prolines in transit peptides are crucial for efficient preprotein translocation into chloroplasts. *Plant physiology* **176**: 663–677.

Li H min, Teng YS. **2013**. Transit peptide design and plastid import regulation. *Trends in Plant Science* **18**: 360–366.

Lin Y-P, Wu M-C, Charng Y-Y. **2016**. Identification of a chlorophyll dephytylase involved in chlorophyll turnover in *Arabidopsis*. *The Plant cell* **28**: 2974–2990.

Lohr M, Schwender J, Polle JEW. **2012**. Isoprenoid biosynthesis in eukaryotic phototrophs: A spotlight on algae. *Plant Science* **185–186**: 9–22.

Lu Y. **2018**. Assembly and transfer of iron–sulfur clusters in the plastid. *Frontiers in Plant Science* **9**: 336.

MacArthur MW, Thornton JM. **1991**. Influence of proline residues on protein conformation. *Journal of Molecular Biology* **218**: 397–412.

Maier UG, Zauner S, Hempel F. **2015**. Protein import into complex plastids: Cellular organization of higher complexity. *European journal of cell biology*.

Marin B, Nowack ECM, Melkonian M. **2005**. A plastid in the making: evidence for a second primary endosymbiosis. *Protist* **156**: 425–32.

Markunas CM, Triemer RE. **2016**. Evolutionary history of the enzymes involved in the Calvin-Benson cycle in euglenids. *Journal of Eukaryotic Microbiology* **63**: 326–339.

Maruyama S, Suzaki T, Weber APM, Archibald JM, Nozaki H. **2011**. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. *BMC evolutionary biology* **11**: 105.

Mateášiková-Kováčová B, Vesteg M, Drahovská H, Záhonová K, Vacula R, Krajčovič J. **2012**. Nucleus-encoded mRNAs for chloroplast proteins GapA, PetA, and PsbO are trans-spliced in the flagellate *Euglena gracilis* irrespective of light and plastid function. *Journal of Eukaryotic Microbiology* **59**: 651–653.

Matson RS, Meifei, Chang SB. **1970**. Comparative studies of biosynthesis of galactolipids in *Euglena gracilis* strain Z . *Plant Physiology* **45**: 531-.

Matsumoto T, Shinozaki F, Chikuni T, Yabuki A, Takishita K, Kawachi M, Nakayama T, Inouye I, Hashimoto T, Inagaki Y. **2011**. Green-colored plastids in the dinoflagellate genus *Lepidodinium* are of core chlorophyte origin. *Protist* **162**: 268–276.

Minge M a, Shalchian-Tabrizi K, Tørresen OK, Takishita K, Probert I, Inagaki Y, Klaveness D, Jakobsen KS. **2010**. A phylogenetic mosaic plastid proteome and unusual plastid-targeting signals in the green-colored dinoflagellate *Lepidodinium chlorophorum*. *BMC evolutionary biology* **10**: 191.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. **2007**. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research* **35**: W182-5.

Moustafa A, Reyes-Prieto A, Bhattacharya D. **2008**. Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions (R DeSalle, Ed.). *PLoS ONE* **3**: e2205.

Muchhal US, Schwartzbach SD. **1994**. Characterization of the unique intron-exon junctions of *Euglena* gene(s) encoding the polyprotein precursor to the light-harvesting chlorophyll a/b binding protein of photosystem II. *Nucleic acids research* **22**: 5737–44.

Nakai M. **2018**. New perspectives on chloroplast protein import. *Plant and Cell Physiology* **59**:

1111–1119.

Nawrocki WJ, Tourasse NJ, Taly A, Rappaport F, Wollman F-A. **2015**. The plastid terminal oxidase: Its elusive function points to multiple contributions to plastid physiology. *Annual Review of Plant Biology* **66**: 49–74.

Nowack ECM, Price DC, Bhattacharya D, Singer A, Melkonian M, Grossman AR. **2016**. Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. *Proceedings of the National Academy of Sciences of the United States of America* **113**: 12214–12219.

O'Neill EC, Trick M, Hill L, Rejzek M, Dusi RG, Hamilton CJ, Zimba P V., Henrissat B, Field RA. **2015**. The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. *Molecular BioSystems* **11**: 2808–2820.

Patron NJ, Waller RF. **2007**. Transit peptide diversity and divergence: A global analysis of plastid targeting signals. *BioEssays : news and reviews in molecular, cellular and developmental biology* **29**: 1048–58.

Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, *et al.* 2019. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research* **47**: D442–D450.

Petersen TN, Brunak S, von Heijne G, Nielsen H. **2011**. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods* **8**: 785–6.

Ponce-Toledo RI, Moreira D, López-García P, Deschamps P, Ruiz-Trillo I. **2018**. Secondary plastids of euglenids and chlorarachniophytes function with a mix of genes of red and green algal ancestry (I Ruiz-Trillo, Ed.). *Molecular Biology and Evolution*.

Ponce-Toledo RI, López-García P, Moreira D. **2019**. Horizontal and endosymbiotic gene transfer in early plastid evolution. *New Phytologist*: nph.15965.

R-Core-Team. **2013**. R: A language and environment for statistical computing.

Reinbothe C, Ortel B, Parthier B, Reinbothe S. **1994**. Cytosolic and plastid forms of 5-

enolpyruvylshikimate-3-phosphate synthase in *Euglena gracilis* are differentially expressed during light-induced chloroplast development. *Molecular & general genetics : MGG* **245**: 616–22.

Saidha T, Na SQ, Li JY, Schiff JA. **1988**. A sulphate metabolizing centre in *Euglena* mitochondria. *The Biochemical journal* **253**: 533–9.

Saidha T, Schiff JA. **1989**. The role of mitochondria in sulfolipid biosynthesis by *Euglena* chloroplasts. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* **1001**: 268–273.

Saldarriaga JF, Taylor FJR, Keeling PJ, Cavalier-Smith T. **2001**. Dinoflagellate nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. *Journal of Molecular Evolution* **53**: 204–213.

Seeger JW, Bentley R. **1991**. Phylloquinone (Vitamin K1) biosynthesis in *Euglena gracilis* strain Z. *Phytochemistry* **30**: 3585–3589.

Sheiner L, Striepen B. **2013**. Protein sorting in complex plastids. *Biochimica et Biophysica Acta - Molecular Cell Research* **1833**: 352–359.

Shibata S, Arimura S, Ishikawa T, Awai K. **2018**. Alterations of membrane lipid content correlated with chloroplast and mitochondria development in *Euglena gracilis*. *Frontiers in Plant Science* **9**: 370.

Sláviková S, Vacula R, Fang Z, Ehara T, Osafune T, Schwartzbach SD. **2005**. Homologous and heterologous reconstitution of Golgi to chloroplast transport and protein import into the complex chloroplasts of *Euglena*. *Journal of cell science* **118**: 1651–1661.

Smith KW, Stroupe ME. **2012**. Mutational analysis of sulfite reductase hemoprotein reveals the mechanism for coordinated electron and proton transfer. *Biochemistry* **51**: 9857–9868.

Soding J, Biegert A, Lupas AN. **2005**. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33**: W244–W248.

Sommer MS, Gould SB, Lehmann P, Gruber A, Przyborski JM, Maier U-G. **2007**. Der1-mediated preprotein import into the periplastid compartment of chromalveolates? *Molecular Biology and Evolution* **24**: 918–928.

Spork S, Hiss J a, Mandel K, Sommer M, Kooij TW a, Chu T, Schneider G, Maier UG,

Przyborski JM. **2009**. An unusual ERAD-like complex is targeted to the apicoplast of *Plasmodium falciparum*. *Eukaryotic cell* **8**: 1134–45.

Stamatakis A. **2006**. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Stiller JW, Schreiber J, Yue J, Guo H, Ding Q, Huang J. **2014**. The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nature Communications* **5**: 5764.

Stork S, Moog D, Przyborski JM, Wilhelmi I, Zauner S, Maier UG. **2012**. Distribution of the SELMA translocon in secondary plastids of red algal origin and predicted uncoupling of ubiquitin-dependent translocation from degradation. *Eukaryotic cell* **11**: 1472–1481.

Sulli C, Fang ZW, Muchhal U, Schwartzbach SD. **1999**. Topology of *Euglena* chloroplast protein precursors within endoplasmic reticulum to Golgi to chloroplast transport vesicles. *Journal of Biological Chemistry* **274**: 457–463.

Takano Y, Hansen G, Fujita D, Horiguchi T. **2008**. Serial replacement of diatom endosymbionts in two freshwater dinoflagellates, *Peridiniopsis spp*. (Peridiniales, Dinophyceae). *Phycologia* **47**: 41–53.

Tanaka Y, Ogawa T, Maruta T, Yoshida Y, Arakawa K, Ishikawa T. **2017**. Glucan synthase-like 2 is indispensable for paramylon synthesis in *Euglena gracilis*. *FEBS Letters* **591**: 1360–1370.

Teng Y-S, Su Y, Chen L-J, Lee YJ, Hwang I, Li H. **2006**. Tic21 is an essential translocon component for protein translocation across the chloroplast inner envelope membrane. *The Plant cell* **18**: 2247–57.

Terashima M, Specht M, Hippler M. **2011**. The chloroplast proteome: a survey from the *Chlamydomonas reinhardtii* perspective with a focus on distinctive features. *Current genetics* **57**: 151–68.

Tessier LH, Keller M, Chan RL, Fournier R, Weil JH, Imbault P. **1991**. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena*. *The EMBO journal* **10**: 2621–5.

Tonkin CJ, Struck NS, Mullin K a, Stimmler LM, McFadden GI. **2006**. Evidence for Golgi-

independent transport from the early secretory pathway to the plastid in malaria parasites. *Molecular microbiology* **61**: 614–30.

Turmel M, Gagnon M-C, O'Kelly CJ, Otis C, Lemieux C. **2009**. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Molecular biology and evolution* **26**: 631–48.

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. **2016**. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13**: 731–740.

Watanabe F, Yoshimura K, Shigeoka S. **2017**. Biochemistry and physiology of vitamins in *Euglena*. In: Schwartzbach SD, Shigeoka S, eds. Euglena: Biochemistry, Cell and Molecular Biology. Cham, Switzerland: Springer International Publishing, 65–90.

Wetherbee R, Jackson CJ, Repetti SI, Clementson LA, Costa JF, van de Meene A, Crawford S, Verbruggen H. **2018**. The golden paradox – a new heterokont lineage with chloroplasts surrounded by two membranes. *Journal of Phycology*: jpy.12822.

Wildner GF, Hauska G. **1974**. Localization of the reaction site of cytochrome 552 in chloroplasts from *Euglena gracilis*: Cytochrome content and photooxidation in different chloroplast preparations. *Archives of Biochemistry and Biophysics* **164**: 127–135.

Xia S, Zhang Q, Zhu H, Cheng Y, Liu G, Hu Z. **2013**. Systematics of a kleptoplastidal dinoflagellate, *Gymnodinium eucyaneum* Hu (Dinophyceae), and its cryptomonad endosymbiont (I Söderhäll, Ed.). *PLoS ONE* **8**: e53820.

Yamaguchi A, Yubuki N, Leander BS. **2012**. Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy to phototrophy: description of *Rapaza viridis* n. gen. et sp. (Euglenozoa, Euglenida). *BMC evolutionary biology* **12**: 29.

Yoon HS, Hackett JD, Van Dolah FM, Nosenko T, Lidie KL, Bhattacharya D. **2005**. Tertiary endosymbiosis driven genome evolution in dinoflagellate algae. *Molecular biology and evolution* **22**: 1299–308.

Yoshida Y, Tomiyama T, Maruta T, Tomita M, Ishikawa T, Arakawa K. **2016**. *De novo* assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. *BMC genomics* **17**: 182.

Záhonová K, Füssy Z, Birčák E, Novák Vanclová AMG, Klimeš V, Vesteg M, Krajčovič J, Oborník M, Eliáš M. **2018**. Peculiar features of the plastids of the colourless alga *Euglena longa* and photosynthetic euglenophytes unveiled by transcriptome analyses. *Scientific Reports* **8**: 17012.

Zhao L, Chang W, Xiao Y, Liu H, Liu P. **2013**. Methylerythritol phosphate pathway of isoprenoid biosynthesis. *Annual Review of Biochemistry* **82**: 497–530.

Ziegler K, Maldener I, Lockau W. **1989**. 5'-monohydroxyphylloquinone as a component of photosystem I. *Zeitschrift für Naturforschung C* **44**: 468–472.

Supporting information:

- Methods S1: Detailed description of plastid fraction preparation and analyses of phylogenetic affiliations and N-terminal domain composition of plastidial proteins.
- Notes S1-S13: Supplementary figures, metabolic schematics, and phylogenetic trees.
- Note S14: Sequence alignment of the novel derlin-like proteins and canonical derlins in fasta format.
- Note S15: Sequence alignments of the respective Suf proteins of algal-like and chlamydial-like origin in fasta format.
- Dataset S1: Full table of annotations for the proteomic dataset, missing enzymes, manually removed sequences, potential dual-localized proteins, and FeS cluster assembly-related proteins.
- Dataset S2: Comparisons of MS-determined plastid proteome with previous predictions.

Tables:

Table 1: Comparison of the proportions of proteins of selected functional categories in the four plastid proteomes; the number in brackets under the respective organism names is the total number of identified plastid proteins.

	E. gracilis (1345)	*A. thaliana* (1462)	*C. reinhardtii* (996)	*B. natans* (324)
Photosynthesis	6.4%	9.1%	11.9%	14.8%
Lipid metabolism	2.9%	4.7%	2.5%	2.2%
Amino acid metabolism	0.7%	5.6%	5.5%	2.2%
Secondary metabolism	2.4%	2.9%	2.7%	2.8%
Tetrapyrrole, cofactor and vitamin metabolism	3.5%	5.3%	3.2%	3.4%
Signalling	3.1%	2.3%	2.6%	3.7%

Figure legends:

Fig. 1: Distribution of functional categories among 775 *E. gracilis* plastid proteins with predicted function (57.5% of the whole proteome) and $\log_{10}$ CP/MT (chloroplast to mitochondrion) ratio distribution in each category, represented by shades of green (0-1 meaning 1-10× higher amount of the protein in plastid fraction compared to the mitochondrial one, 1-2 for 10-100×, 2-3 for 100-1000×; and >3 for more than 1000× or "infinite" value in case the protein was detected in plastid fraction only; the full category names are listed in the Materials and Methods section, for detailed description with examples see Supporting Information Notes S3).

Fig. 2: Overview of the *E. gracilis* plastid metabolism as reconstructed from mass spectrometry-based proteome. Enzymes present in the plastid proteome in at least one isoform are marked as green circles, grey circles represent enzymes which were identified on the RNA or DNA level (in this study or previously) but are absent from the proteome; white circles represent genes completely absent in *Euglena*; circles marked by the letter P represent genes coded in the plastid genome while the rest of the circles represent genes coded in the nucleus. Circles with colored dots in the middle represent genes with at least one of their isoforms presumably gained via lateral transfer from a donor group other than green algae and "miscellaneous" algae: pale green for chlorarachniophytes, brown for ochrophytes, dark orange for haptophytes, pale orange for cryptophytes, red for rhodophytes, and a combination of the former in case of proteins of "mixed red" origin (see Supporting Information Notes S4 for larger and colorblind-friendly version). Multiple overlapping circles represent enzymes composed of multiple subunits with different characteristics. Note that a single protein may be represented by multiple circles due to its role in multiple pathways or reactions. A more detailed schematic including additional pathways and transcript identifiers for each enzyme is available as Notes S5.1-11.

Fig. 3: Reconstruction of *E. gracilis* plastid protein import machinery: The N-terminal signal-peptide (SP, red) bearing nuclear-encoded plastid-targeted protein (blue chain) is synthesized on the rough endoplasmic reticulum (RER). If the N-terminus does not contain stop transfer signal (STS), it is co-translationally imported into the ER lumen as indicated. Otherwise, the major part of the protein remains in the cytosol while being anchored in the ER membrane. The SP is cleaved by signal peptidase in the ER lumen. The protein then passes through Golgi and is loaded

into/onto a vesicle which eventually fuses with the outermost plastid membrane. Plastidial homologs of GOSR and Rab5 GTPase might mediate the fusion. The protein passes the middle membrane via unknown mechanism dependent on transit peptide (TP, purple) but not TOC complex, possibly employing derlin-like proteins (DerL). Then it passes the inner membrane via highly reduced TIC translocase, possibly consisting of multiple isoforms of a single subunit. The protein is folded and has its transit peptide cleaved in the plastid stroma and, in case additional signal is revealed, enter thylakoid membrane or lumen via TAT, SEC, or SRP/Alb3 pathway. Each circle represents a putative subunit described in model plastids, green circles represent proteins identified in the plastid proteome, different color shades code CP/MT ratio, the darkest shade depicting the most credible evidence for plastid localization, grey circles represent subunits with transcript-level evidence only. Finally, white circles represent subunits which are completely absent.

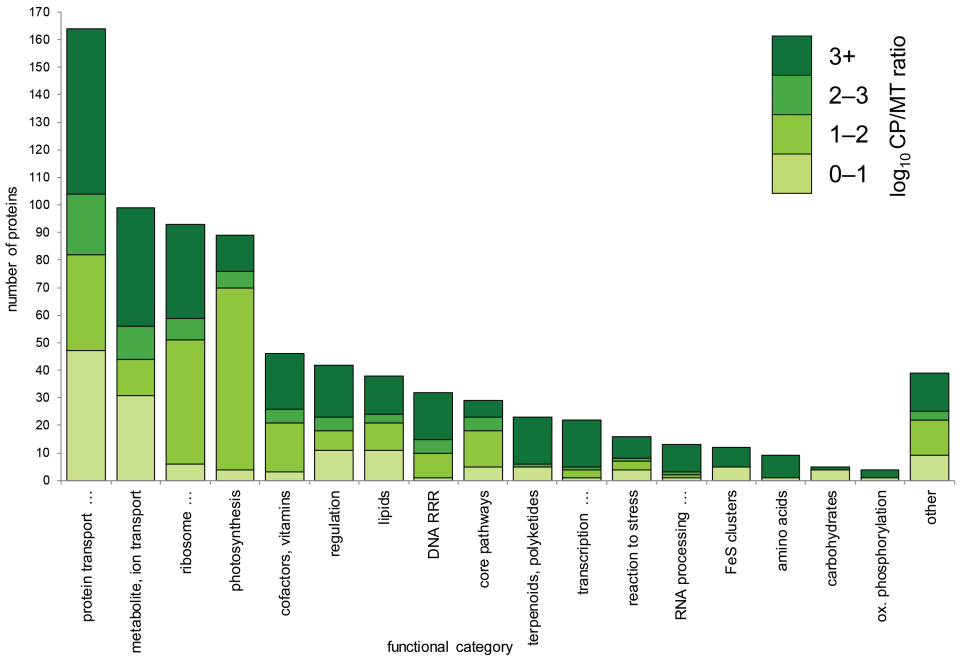
Fig. 4: Evolutionary affiliations of 416 nuclear-encoded plastid proteins presumably representing lateral gene transfer into or from E. gracilis. Protein trees are available at http://www.protistologie.cz/hampllab/data/Novak_Vanclova_trees.zip.

Fig. 5: Plastidal SUF pathway of *E. gracilis*. Multiple copies of each subunit are represented by circles colored in shades of blue based on their average label-free quantification (LFQ) intensity per replicate which indirectly corresponds to the relative protein abundance. Proteins captured in only one replicate are also included. The evolutionary origin of the particular subunit copy is represented by the dot in the middle: green for green algae-related, i.e. gained along with the plastid via endosymbiotic gene transfer (EGT), or magenta for chlamydial-like, i.e. gained via lateral gene transfer (LGT). The presence of predicted signal and transit peptides is also indicated.

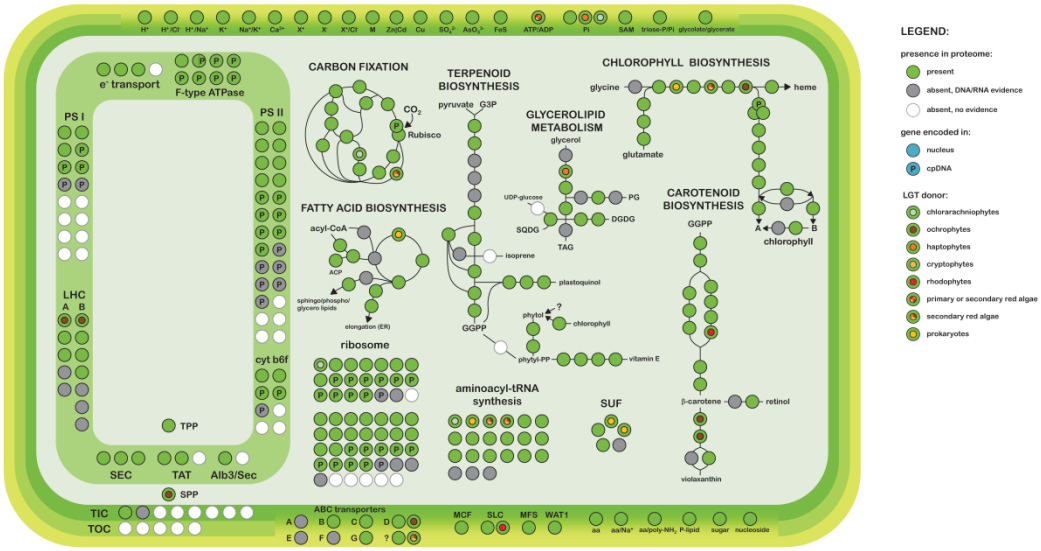
Fig. 6: The two hydrophobic domains in the N-termini of 375 well-supported *E. gracilis* plastid-targeted proteins with complete N-terminus as seen on the graph of average hydrophobicity score per position compared to the negative control (375 randomly selected proteins of *E. gracilis*). The range of the signal peptide (SP) was estimated as positions 9-60 (note, that the first and last six positions of the 160 aa long sequence are missing as a result of window-based scoring; position 7 is the first position with score assigned), the range of the stop-transfer signal (STS) was estimated as positions 97-130. The hydrophobic peak representing the STS is clearly less accentuated than

the one representing SP reflecting the fact that it is present in just a subset of plastid-targeted proteins termed class II pre-proteins.

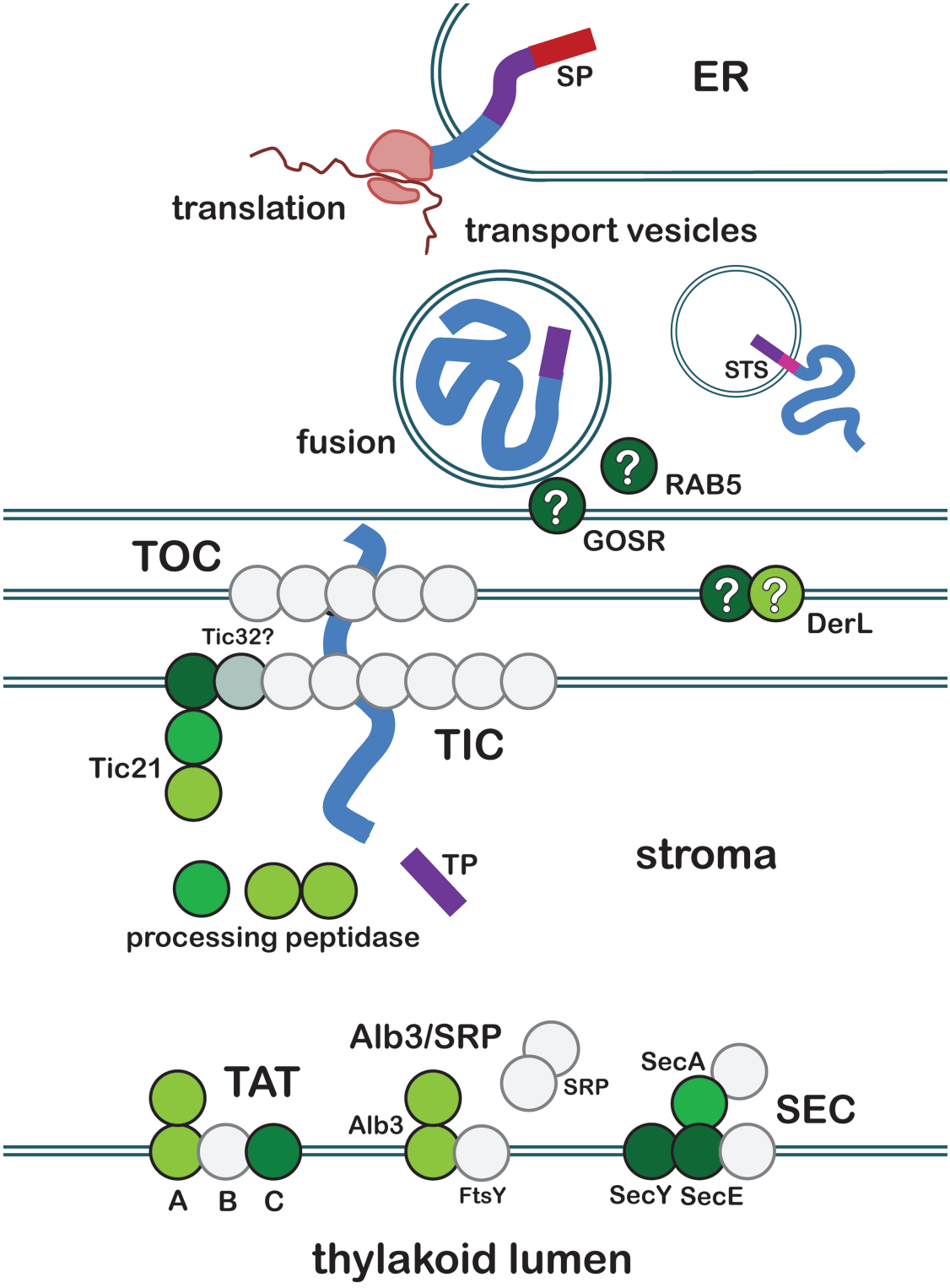
Fig. 7: Statistical comparison of the overall amino acid frequencies in the putative transit peptide region and the putative mature chain of the same protein for the set of 375 *E. gracilis* proteins regardless of their classification for the plastid protein sample in comparison to the negative control (shades of blue and orange, respectively, a). The same comparison was also performed separately for the plastid proteins of class I, class II, and "unclassified" (shades of red, orange, and yellow, respectively, b). The vertical axis represents normalized χ^2 sum reflecting whether the amino acid frequency is higher or lower than expected (positive or negative values) and the relative degree of its enrichment or depletion. The medium coloured bars represent statistically significant results with $p < 0.01$, the dark coloured bars represent those with $p < 10^{-5}$.



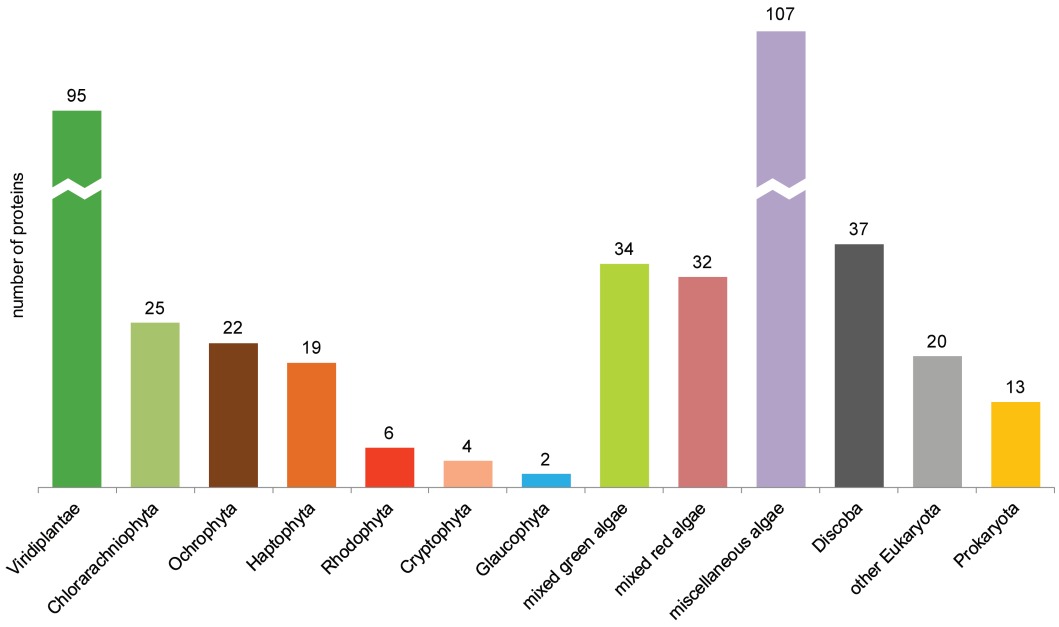
nph_16237_f1.tif



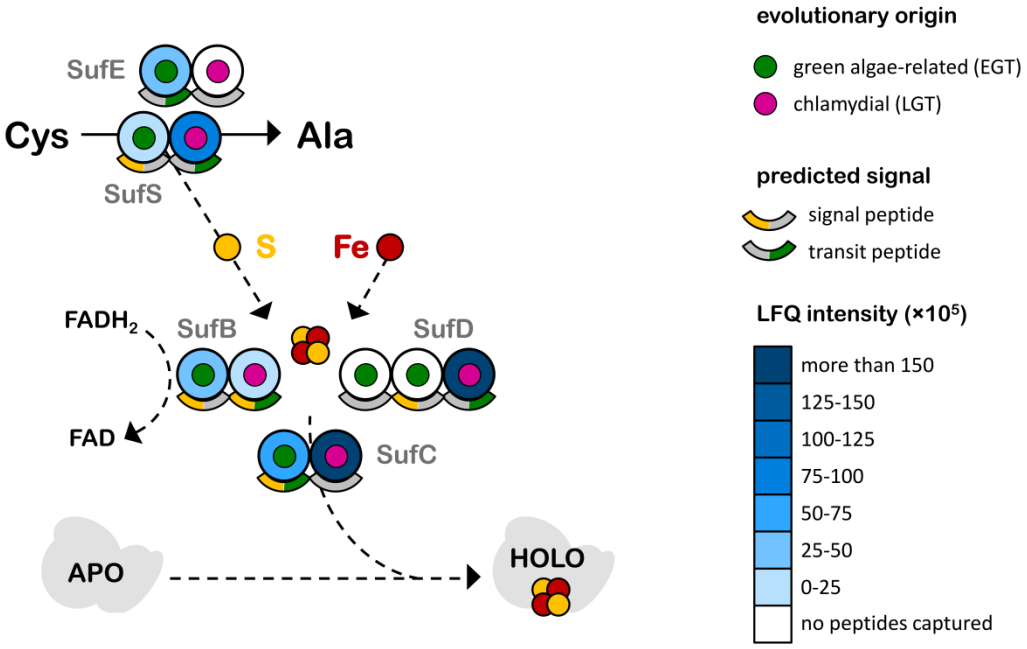
nph_16237_f2.tif



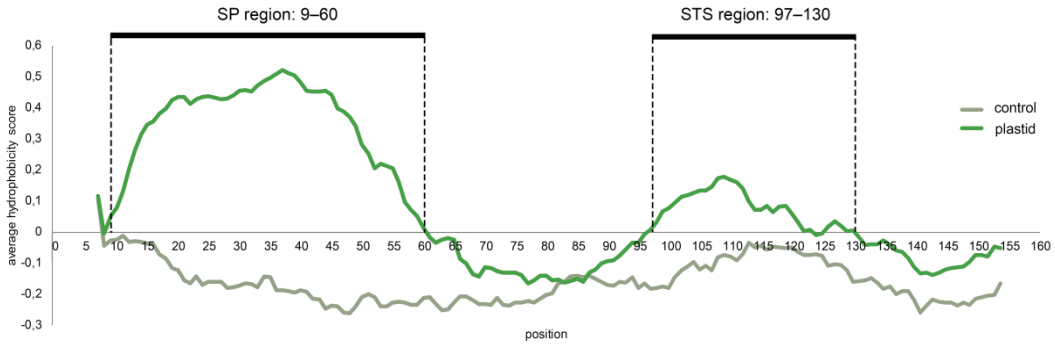
nph_16237_f3.tif



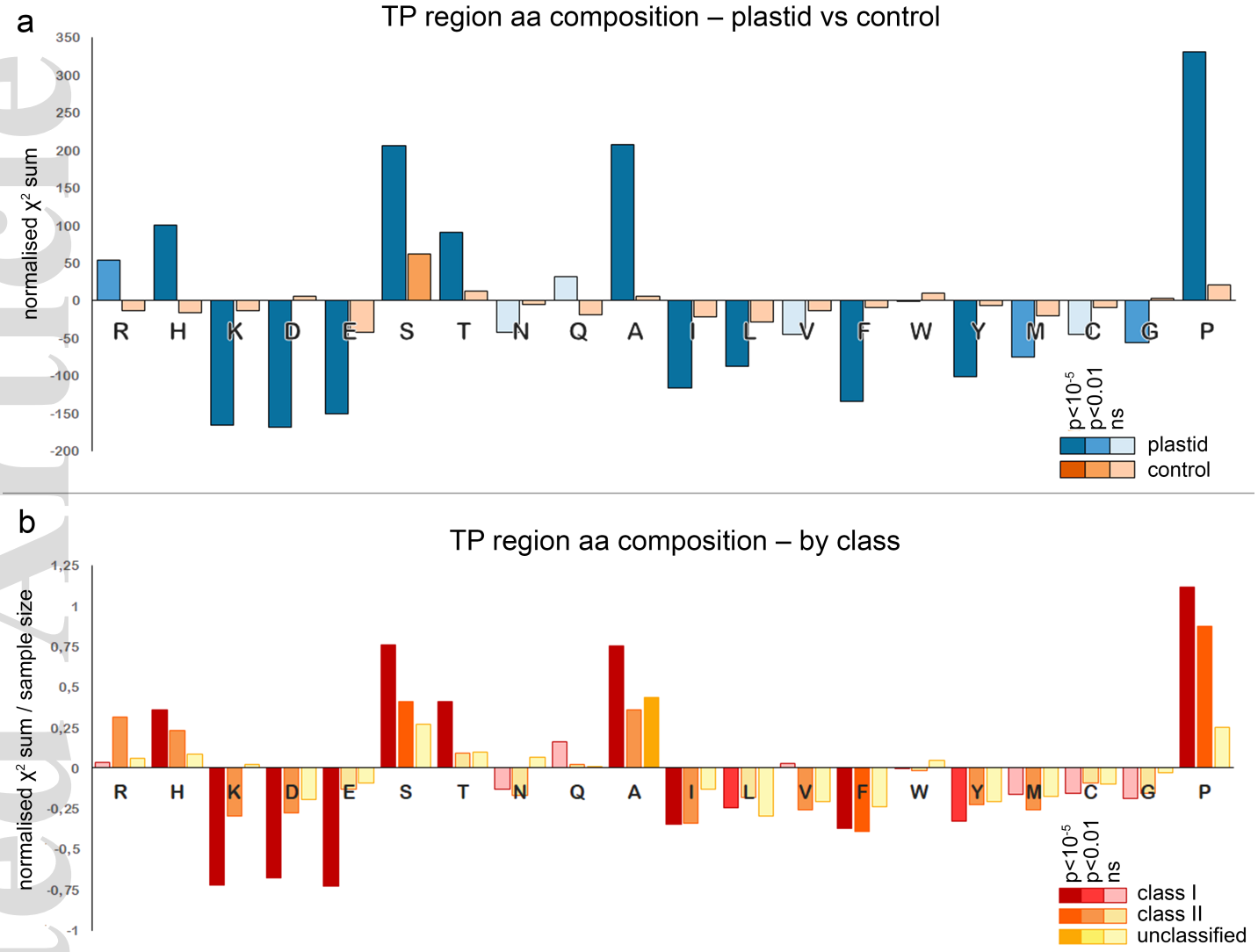
nph_16237_f4.tif



nph_16237_f5.tif



nph_16237_f6.tif



nph_16237_f7.tif